基于人工智能的检校系统应用及探索

摘　要：随着媒体融合进入深水区，新闻内容来源更多、数量更大，传播渠道也更加多元化，时效性要求也越来越高，媒体机构内容生产的质量管控和发稿安全面临新的挑战。目前，将人工智能技术应用在校对中，已能实现判断文稿内用词、用句的准确性和合理性，甚至能分辨出感情色彩矛盾或者逻辑不通的地方。新华报业传媒集团全媒体指挥中心采用全流程智能检校和不同系统的交叉检校，贯穿内容生产的各个关键环节，并利用检校结果深度学习，形成不断迭代完善的闭环。

关键词：检校；校对；审核；人工智能；NLP；内容安全
中图分类号：TP18　　文献标识码：A
文章编号：1671-0134（2019）10-120-03　　DOI：10.19483/j.cnki.11-4653/n.2019.10.037

文／杨更修　孙甲飞　冯恩达　殷琳

1. 项目背景

新闻稿件的文字校对，是新闻生产发布过程中必不可少的重要环节，是保障发稿安全、维护新闻严谨性的关键防线。各大媒体出版机构的稿件审核流程虽不尽相同，但在正式发布之前各家都有一个相同的环节——校对。

伴随着不同时期媒体行业对文字校对的需求，校对系统先后经历了三代的发展：第一代系统主要基于计算机的存储和基本运算，通过积累大量的错词库，对稿件的文字内容进行逐字、逐词匹配，实现词汇级的检校；第二代系统采用智能技术来实现整句级别的文字检查，能够根据句子整体表达的语境，识别其中词汇的不合理搭配问题；第三代检校系统是一种类人系统，在第二代系统的能力基础之上，通过深度学习实现语义分析，对稿件内容进行全面分析和理解。在把握全文的观点、基调的基础上，判断文稿内每句话、每个字词是否合理，是否存在感情色彩矛盾或者逻辑不通顺的地方。

随着媒体融合进入深水区，新闻的传播渠道也越来越多元化，时效性要求也越来越高，市场对内容生产的速度、广度、深度、总量都提出了更高的要求，媒体机构内容生产的质量管控和发稿安全面临新的挑战。全媒体指挥中心项目利用当下语义分析和深度学习的最新发展成果，在内容生产流程中探索引入人工智能检校，并对检校效果进行统计评估。

2. 智能检校技术分析

2.1 智能检校的难点

智能检校的难点在于对情感和语义的分析，在全文的基调上，判断每个词、每句话是否合理。目前主流的文章情感分析包括基于情感词典的分析和基于机器学习的分析。

大多数的文章情感分析主要是针对学习词典的建模分析和机器学习算法进行研究，通过对情感词典、否定词词典、程度副词词典、停用词词典分析，计算上下文情感倾向的方法。分析新闻主题和词语修饰成分之间的搭配关系来计算词语极性，综合词典资源用于构建情感词库，同时采用加权线性组合方法，以实现判断文章的情感倾向。

基于机器学习的文章情感分析方法是将情感视作一种多分类问题，属于有监督的学习方法。机器学习方法要经过文本的预处理、特征选择、特征加权、训练分类器并进行分类等过程。该方法的分类性能要优于传统的特征加权方法 TF-IDF（term frequency – inverse document frequency）。

2.2 自然语言处理的应用

自然语言处理（Natural Language Processing）是信息时代最重要的技术之一，是人工智能的重要组成部分。基于 NLP 技术衍生出的应用已经在各领域得到广泛运用，包括拼写检查、机器翻译、语音识别、聊天机器人等。

深度学习提供了一个灵活、通用、可学习的框架，它在语音识别和计算机视觉领域取得了突破性的进展。检校工作主要是跟文字相关，NLP 可以让计算机实现对内容的阅读和理解，对错误处给出提示，实现检校工作的自动化。

2.3 智能检校系统的构建

针对目前主流的检校系统，通过搜集大量错误录入字词的典型可以发现，中文检校系统最常见的错误包括字词级错误、语法级错误和语义级错误。字词级错误主要由错字、别字、少字、多字、异位引起。通过对稿件的文字内容进行逐字、逐词匹配，将与错词库中内容相匹配的词认定为字词错误，提示给使用者。比如：“倡仪”（倡议）、“国冢”（国家）、“总埋”（总理）；语法级错误主要指词语的错误搭配或者漏字等情况。通过大量学习正确语料，让计算机系统自主分析归纳语言的习惯用法、模式等，使机器对句子形成一定的理解和判断能力，从而实现在一个句子的维度上对字、词进行分析判断，识别其中的异常、不合理内容，达到检查校对的目的。

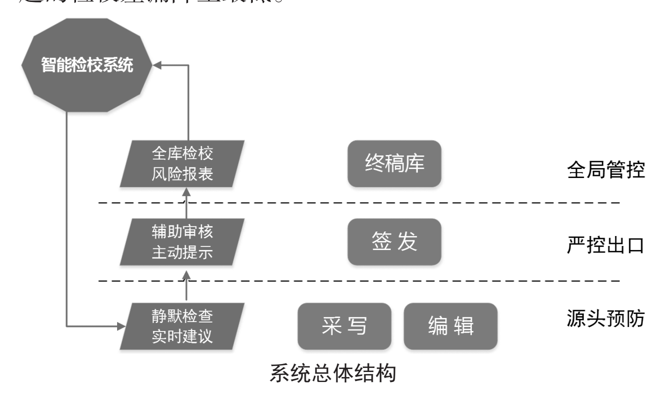
智能检校系统在全媒体指挥中心的应用不仅实现了词汇检查、语句检查，还能对情感做一定分析，对稿件内容进行全面分析和理解。在全文观点、基调的基础上，判断每句话、每个字词是否合理，是否存在观点矛盾或者逻辑不通顺的地方。通过基于主题融合的深度学习，用中文文本预处理方法将非结构化或半结构化的信息转换为计算机能理解的结构化信息，对内容进行全面分析和理解，从而自动识别文本的情感类别，实现校验的智能化。

文章的主题与文章情感通常存在一定的共性，深度学习模型可以通过融合向量提高文章情感分类模型的准确率。检校系统引入双向 LSTM 情感算法，实现词语的上下文信息融合，既克服了传统 RNN 的梯度消失问题，还解决了传统 LSTM 只能较好地融合上文信息、缺少下文信息融合的问题。通过融合文本的主题特征，构建更精准的情感分类模型。

3. 构建全流程的内容安全

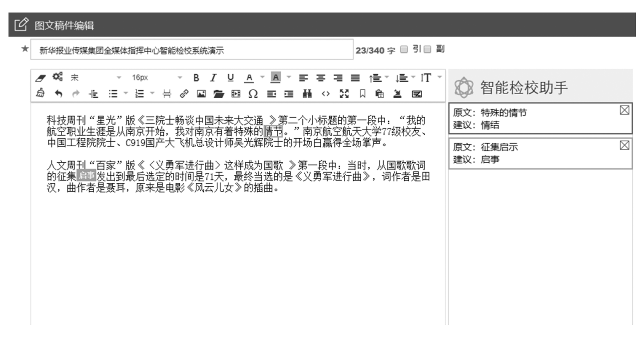
传统的新闻稿检校工作通常是稿件发布前的最后一个环节，检校的时间紧、任务重，检出的错误需要退回修改然后再检校。而在媒体深度融合大环境下，稿件数量井喷式增加，移动端的实时新闻经常追求最快速度发稿，晚一秒钟可能就失去了这条新闻最佳的传播机会。在这样的环境下，再把校对工作完全放在发布前的最后一个环节的做法，在实践中已经不能满足现今媒体新闻稿件多形式、低时间宽容度、零错误容忍度的要求，更难以满足未来建设“四全媒体”的长远目标。智能检校系统将主动检校和自动检校结合起来，采用 SAAS 布局模型，使智能检校系统既可以嵌入稿件编辑系统又能作为独立的辅助审核模块使用。

智能检校工作分布在内容生产的各个关键环节，编辑随时都可以将当前编写的稿件内容发起人工智能检校。这样就将查错、纠错的时间分摊到稿件流转的过程中，减轻最后检校环节的压力，将因时间过紧和数量过多引起的检校差漏降至最低。



系统总体结构

在稿件采编环节，检校系统实时参与其中，编辑记者可以点选检校，系统会对文字稿件进行词语错用、语义表述错误等提示并给出修改建议，为编辑写稿把好第一关。同时，通过检校智能助手与编辑进行互动，编辑点击右侧的每条提示，编辑框中的焦点会随之定位，节省了编辑再去原文中找对应点的时间。与此同时，在编辑对所提示错误做出修改或忽略的决策时，智能检校系统会对这一决策进行记录与学习。



采编环节检校

在稿件签发环节，如果编辑在提交新闻稿件时没有将稿件中的问题完全修改完善，或者是修改后又引发了新的错误，编辑没有注意直接提交至了稿库。在该稿件签发时，审核人员可以利用智能检校系统会再一次对稿

件进行重新检校。通过在流程必经节点上实施二次检校，尽早将差错的纠正工作往流程的前面节点安排。



签发环节检校

将智能检校的环节前置并不意味着在稿件发布之前不再进行检校，稿件进入签发库后还会进行全库检查。为了避免同一个智能检校系统存在检校结果上的趋同性定势，全媒体指挥中心系统引入另一套检校系统对“终稿库”的稿件进行批量检校，并给出错误风险提示。

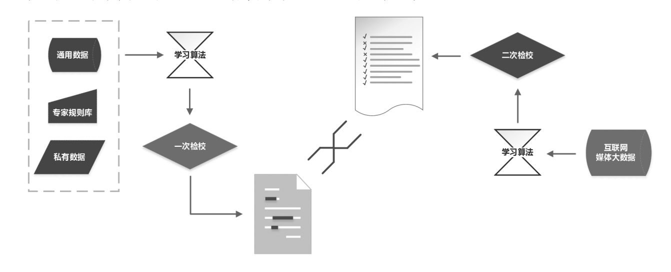
验证文章

序号	文章标题	错误个数	操作
1	南京“社治会”八百公益项目寻“买家”	4	详情
2	常州地铁1号线开通	0	详情
3	重庆大学建筑品博物馆？已暂停开放 三大疑问待回应	2	详情
4	世界粮食日|保护粮食安全，新沂检察在行动	0	详情
5	南京六合调整住房限购政策 外地大专以上人才无需提供社保证明	0	详情
6	IMF下调今明两年世界经济增长预期	3	详情
7	《求是》杂志发表习近平总书记重要文章	0	详情
8	高中生坠亡电竞酒店：失联月余后联系父亲，“他们就先要钱”	9	详情
9	未来三天，江苏大部分地区多云，最低温10℃	3	详情
10	人民日报：发挥制度建设的中国智慧	0	详情
11	全省设区市“不忘初心、牢记使命”主题教育推进会召开	1	详情
12	在黄河流域生态保护和高质量发展座谈会上的讲话	0	详情
13	做好前后衔接 加强上下互动	2	详情
14	金融反腐力度持续加大	1	详情
15	赣榆 | 多措并举扎实开展主题教育活动	4	详情
16	习近平应约同法国总统马克龙通电话	2	详情
17	巡出真问题 增强群众获得感	16	详情
18	立足职责盯紧抓实专项整治	3	详情

批量交叉检校错误风险提示列表

4. 双系统交叉检校

目前，单个基于语义分析与深度学习的智能检校系统在现实应用中还会出现一些未能检测出的错误，基于不同的语料库的学习结果也会出现对词汇、语义、情感等元素理解判断上的差异。智能检校系统除了将检校工作在流程中分层前移以外，同时引入了两套不同的智能检校系统，利用两套系统对新闻稿件进行交叉检校。第一套系统负责对单个稿件进行检校，第二套系统负责将通过第一套系统检校过的稿件再一次全文检校，并通过统计列表将签发库中稿件的问题形成差错警示表，并将此结果反馈给智能检校系统的学习模块，使系统不断自我完善。如此一来，就可以充分利用各家所长，最大限度提升智能检校对稿件质量的把控效果。



结语

结合智能检校系统的应用，通过对内容生产流程进行融合再造，将自然语义分析与深度学习的技术成果引入内容生产全流程。经过一段时间的运行，从采编人员的使用情况调研和每阶段的稿件差错统计报告来看，比传统检校更有优势，检出了一些传统检校不能检出的关键错误。

未来，智能检校系统将进一步在基于私有数据学习和基于互联网大数据学习两个方面不断完善，通过本地化学习进一步完善检校规则，不断增强其严谨性；通过互联网大数据学习，跟进行业龙头在稿件检校标准方面的发展，同时及时了解互联网新生表达方式，充分发挥出全流程检校和交叉检校的叠加作用，达到“1 + 1 > 2”的效果。媒

（作者单位：新华报业传媒集团）